DOCUMENT RESUME

ED 246 104 TM 840 379

AUTHOR Melican, Gerald; Thomas, Nancy
TITLE Identification of Items That Are Hard to Rate Accurately Using Angoff's Standard Setting Method.
PUB DATE Apr 84
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Cutting Scores; *Difficulty Level; Minimum Competency Testing; Research Methodology; *Test Items
IDENTIFIERS *Angoff Methods

ABSTRACT

Setting standards for the purpose of certification is frequently performed using judgmental techniques such as the Angoff method. This study was performed to identify types of items that judges find hard to rate accurately, that is, types of items on which examinees perform differently than predicted by the judges. Once identified these item types should be the subject of additional training for future panels of judges. Observed difficulty estimates were regressed on the Angoff estimates of a panel of judges for three examinations and items with large residuals were identified. For one examination there were indications that the judges found negatively phrased items hard to rate accurately. For a second examination there were indications that the judges found items involving calculations as hard to rate accurately. In both cases the judges tended to underestimate the difficulty of the items; the examinees found the items harder than the judges predicted. No discernible pattern was found for the third examination. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246104

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Identification of Items That Are Hard to Rate Accurately Using Angoff's Standard Setting Method

Gerald Melican and Nancy Thomas

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. L. Melican

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the

American Educational Research Association

New Orleans, April 1984

TM 840 375

ABSTRACT

Setting standards for the purpose of certification is frequently performed using judgmental techniques such as the Angoff method. This study was performed to identify types of items that judges find hard to rate accurately, that is, types of items on which examinees perform differently than predicted by the judges. Once identified these item types should be the subject of additional training for future panels of judges. Observed difficulty estimates were regressed on the Angoff estimates of a panel of judges for three examinations and items with large residuals were identified. For one examination there were indications that the judges found negatively phrased items as hard to rate accurately. For a second examination there were indications that the judges found items involving calculations as hard to rate accurately. In both cases the judges tended to underestimate the difficulty of the items, the examinees found the items harder than the judges predicted. No discernable pattern was found for the third examination.

INTRODUCTION

Setting a standard for an examination for the purpose of certification or licensure is most frequently performed using judgmental techniques such as the Angoff or the Nedelsky methods. It is crucial that the judges be subject experts who are aware of the level of knowledge required of the "minimally competent" examinee. The judges should be given as much training and information as possible to assure that their judgments are objective and reliable.

The Angoff method (Angoff 1971) for establishing a cutoff score requires a panel of expert judges to estimate the proportion of minimally competent examinees who will answer each item of an examination correctly. This assumes that, once a definition of minimal competency is established, each judge can determine the difficulty of each item for a group of minimally competent examinees. There should be a strong linear relationship between the judges estimates and the observed difficulties for the group of minimally competent examinees as defined by the judges.

One possible cause for departure from linearity is the judges' inability to rate certain types of items as well as they rate other types of items. Bejar (1981) analyzed the abilities of expert judges to estimate the difficulty of items for the total group of examinees from the Test of Standard Written English (TSWE). The TSWE included two item types, Usage items and Sentence Correction items. As one result of the study Bejar reports that the judges were better able to predict the difficulty of the Sentence Correction items.

Item types that are especially hard for judges to rate should be identified and analyzed so they can be given extra attention or training the next time the examination is reviewed. The purpose of this study was to analyze the results of three standard setting studies of certification examinations to ascertain whether an item type or item types that proved especially hard to rate accurately could be identified.

REVIEW OF THE LITERATURE

Several studies have been concerned with the relationship between judges' perceptions of the difficulty of items and the actual difficulty of these items. Lorge and Kruglov (1953) analyzed two groups of judges under two conditions, knowledge and no knowledge. Under the knowledge condition the judges were aware of the actual difficulty of 10 of the 45 items they were to rate. One group had correlations of .84 and .87 for the no information condition and information condition, respectively. The corresponding correlations were .80 and .85 for the second group. They concluded that estimates of item difficulty will be better when judges are given basic information.

In a similar study Thorndike (1980) spread items with known difficulty among items that had not been administered. Thorndike correlated only the items with known difficulties obtaining results similar to Lorge and Kruglov. Thornikes correlations ranged between .72 and .83. Noting that both Lorge and Thorndike had used quantitative examinations Bejar (1981) asked subject matter experts to estimate the difficulty and discrimination indices of items from the TSWE. Bejar reported correlations between estimates of difficulty and observed difficulties ranging from .31 to .43 before discussion and between .35 and .47 after discussion. In a second study Bejar provided extra training to the experts but reported correlations that were disappointingly low. For Usage items the correlation between estimated and observed difficulties was .46 but for Sentence Completion items the correlation was only .21. The correlation was .39 for both types of items combined.

Bernknopf, Curry, and Bashaw (1979) reported correlations of .74 and .85 between the judges' estimates and observed difficulties for the total group using a modified Angoff method. The examination involved was a school counselor certification exam.

These studies suggest that judges may provide reasonably accurate estimates of difficulty and that additional information improves their ability

slightly. This study is concerned with identifying items that are hard to rate accurately so that extra information can be supplied for these types of items.

METHOD

The Angoff method for establishing a cutoff score was utilized with three examinations for certification in a business management field. The three examinations, given under the auspices of the American Production and Inventory Control Society, were: the Materials Requirement Planning (MRP) examination, the Master Planning (MP) examination, and the Inventory Management (IM) examination. There are several dimensions of possible interest represented in each of the three exams. One dimension is calculation/no calculation, another is graphical/non graphical material, and a third is cognitive level with items ranging from knowledge to evaluation. There are several item types included in all three exams. Two of the examinations (MP and MRP) consisted of 100 items and the third (IM) contained 110. Two cutoff scores are developed for each of the three examinations: one to identify entry level practictioners and a higher cutoff score to identify advanced level practitioners. The advanced level cutoff score was chosen fo analysis because the judges were more familiar with the the level of expertise required for advanced practitioners. There was a break between the rating sessions.

The Angoff method requires a panel of judges to rate each item of an examination, specifying the probability that a minimally competent examinee has of answering the item correctly. The cutoff score for the examination is calculated for each judge by summing these probabilities over all items. The average over all judges is generally taken as the cutoff score as suggested by the panel.

Three panels of six expert judges were convened. The meetings began with a detailed presentation of the rationale for performing a cut score study. The decisions to be made based upon the cut scores were emphasized during the discussion. A prolonged discussion of the qualifications of a minimally competent candidate followed. The qualifications and characteristics of such a candidate were enumerated and discussed by the judges. The first item in each examination was rated independently by each judge and then discussed by all. The second item was rated independently and discussed and then items 3 through 10 were rated independently before a discussion was held. Items for which there were disparate ratings were discussed allowing the judges to derive a common definition of minimal competency. Each judge independently rated each item, specifying the proportion of minimally qualified candidates who would answer each item correctly. The judges reconvened and discussed their responses. In this manner the judges are afforded the opportunity to reconsider their ratings, if necessary, and check that they had recorded their ratings accurately. Concensus was not required.

Reliabilities and rater-total group correlations were computed. The rater-total group correlations were corrected for spuriousness by excluding the respective rater from the total group. The proportions reported by the judges were summed over all items arriving at a suggested cutoff score for each judge and these were then averaged over all judges arriving at a suggested cutoff score for the panels. The examinations were then administered and the examinees with scores close to (both above and below) suggested cutoff score were identified. The band of scores defining "close to" was defined using the two cutoff scores developed by the panels. The distance between the upper and lower cutoff scores were taken into account. The proportion of these examinees who answered each item correctly was computed and an arcsine transformation was performed. The transformed observed difficulties were regressed on the transformed estimated difficulties and the residuals were calculated. Large residuals would indicate items that were hard to predict. All items with residuals greater in absolute value than 1.65 were identified and reviewed.

RESULTS

Judges . The cutoff scores suggested by the panels were 84,79, and 85 for MRP,MP, and IM, respectively. The cutoff scores suggested by each judge and related statistics are reported in Table 1. Using the two cutoffs suggested by the judges (one for entry level and one for Fellow level) the range of scores for minimal competency were identified. Examinees with scores within four points of the cutoff score for MRP and MP and examinees within three points of the cutoff score for IM were identified as the subjects for the study.

TABLE 1

Cutoff Scores Suggested by Each Judge:

Means, Standard Deviations, and Reliabilities

Judge*	MRP	MP	IM
1	88.23	78.30	89.81
2	84.34	76.55	88.76
3	84.48	77.20	83.86
4	81.59	79.60	76.25
5	82.25	82.00	85.25
6	84.26	82.80	89.00
Mean	84.19	79.41	85.49
Standard Dev.	2.09	2.33	4.65
Reliability	.62	.66	.72

* The Judges were different for each study. Thus "Judge 1" represents a different person in each column.

The rater-total group correlations ranged from .21 to .42, from .35 to .55, and from .26 to .55, respectively. The correlations of each judge with the average of the remaining judges are reported in Table 2. These correlations suggest that the MRP judges tended to disagree more than the judges of the other two examinations.

TABLE 2

Correlations of Each Judge With the Average of All Other Judges

Judge	MRP	MP	IM
1	.40	.50	.37
2	.21	.39	.50
3	.42	.43	.44
4	.33	.26	.54
5	.29	.52	.59
6	.32	.55	.49

Examinees The means, standard deviations, number of examinees, and reliabilities are reported in Table 3.

TABLE 3

Summary Statistics for Each of the Three Examinations, MRP, MP, and IM.

	MRP	MP	IM
Mean	59.34	56.96	64.74
St. Dev.	14.06	12.47	11.51
Reliability (KR-20)	.90	.88	.85
N of Total Cases	2525	995	1505
N of "Minimally Competent" Cases	116	86	197

Regression Results The correlations between the estimated and the observed difficulties were .50, .54, and .56, for MRP, MP, and IM, respectively. When these correlations were adjusted for unreliability the correlations were .67, .71, and .72. There were nine aberrant items (items with residuals greater in

absolute value than 1.65) for all three examinations. Four of the nine (44%) aberrant items in the MP examination involved negative instructions (e.g. choose the response that is NOT correct). There were only 25 (25%) of the items in the total test that included negative instructions. The probability of finding four or more items of this type in a sample of nine is .16, not generally considered statistically significant but suggesting that the judges did find it more difficult to rate negatively phrased items than other types. In all but one case the judges underestimated the difficulty of the item , that is, the items were more difficult for the examinees than was predicted.

Six of the nine items (67%) for the IM examination involved calculations when for the total test of 110 items only 35 (32%) involved calculations. The probability of finding six or more items of this type in a sample of nine is less than .01. The judges tended to underestimate the difficulty of the calculation items for examinees at the advanced level.

There was no discernible pattern for the nine aberrant items in the MRP examination.

DISCUSSION

The results suggest that there may be item types that are harder for some panels to rate and that these types can be identified via regression analyses. For two of the three cutoff score studies item types that were hard to rate accurately were identified. In one case, the MP examination, the item type identified involved negative instructions. The other two examinations did include negative instruction items but to a much lesser degree. The MRP examination included seven and the IM examination included six. In neither MRP or IM was a negatively stated item identified as aberrant.

In the other case, the IM examination, the item type identified involved calculations. The other examinations did include calculation items, MRP had nine, none of which were identified as aberrant. The remaining examination, MP, had 24 of which only one was identified as aberrant.

The judges tended to underestimate the difficulty of the items identified as aberrant, the examinees found the items more difficult than predicted by the judges. Techniques to decrease the discrepancies between the judges estimates and the observed difficulties should be identified and used. Lorge and Kruglov (1953) reported that supplying judges with empirical information regarding a sample of items decreases the tendency to overestimate or underestimate the difficulty of items. This technique, applied particularly to items that previous judges have found difficult to rate, should increase the accuracy of a future panel of judges. The test development department is continuing to analyze the types of items that proved hard to rate accurately and will be developing additional training information to use during the next cycle. Although the possibility exists that these results are due to the choice of experts or others sources of error, this research study indicates that item types that are hard to rate accurately may be identifiable. The most apparent result of this study, however, is the identification of areas in need of further research.

FURTHER RESEARCH

There are at least three areas which need further research. The first area involves the identification of the factors that make certain item types hard to rate accurately.

The choice of judges could obviously have affected the results, another panel may not have found the same item types hard to rate. A study in which two panels perform a cut score study on the same data at another time would address this point. The composition of the panel; item writers, practitioners, etc., is another part of this research topic.

The choice of techniques (Angoff or Nedelsky, for example) will probably affect which item types are identified as hard to rate accurately. Having the same panel use two different techniques, counterbalancing for order would address this point.

The content of the examination may be another factor influencing the ratings, for example, negatively phrased items may not be difficult to rate in a nursing examination. Replicating the present study with examinations of different content areas would address this issue.

The second area of research involves identification of techniques to improve the training of the judges.

Supplying the judges with additional information about item types that a previous panel found hard to rate accurately may improve their ability to rate the same type of items. A study in which some judges are trained using empirical information with no additional emphasis on the hard to rate types while other judges review the same information and additional information regarding the hard to rate accurately item types.

The types of information that would improve the training of judges need to be identified. Providing judges with empirical information represents a start, but other types of information, such as a discussion of the reasons the examinees found the items more (or less) difficult than predicted need to be researched.

Providing judges with information from a previous administration of an examination obviously requires that the examination be administered. What can be done to improve the judges' abilities without this information is an important area of research. The identification of item types that are consistently found to be hard to rate accurately is the logical first step in addressing this issue.

REFERENCES

Angoff, W.H. Scales, Norms, and Equivalent Scores. In R.L. Thorndike (ed.), Educational Measurement. Washington, D.C., American Council on Education, 1971, pp. 514-515.

Bejar, I.I. Subject Matter Experts' Assessment of Item Statistics (RR-81-47). Princeton, N.J.: Educational Testing Service, 1981.

Bernknopf, S., Curry, A., and Bashaw, W.L. A defensible model for determining a minimal cut-off score for criterion-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

Lorge, I. and Kruglov, L.K. The Improvement of the Estimates of Test Difficulty. Educational and Psychological Measurement, 1953, 13, 34-46.

Thorndike, R. L. Item and score conversion by pooled judgment. Educational Testing Service Research Conference on Test Equating. Princeton N.J., 1980.